NINTH LECTURE SUMMARY

REGRESSION AND CORRELATION

HERE, WE STUDY RELATIONSHIP BETWEEN TWO VARIABLES (EACH ON AN INTERVAL SCALE) USUALLY CALLED *X* (EXPLANATORY) AND *Y* (RESPONSE).

ASSUMPTIONS:

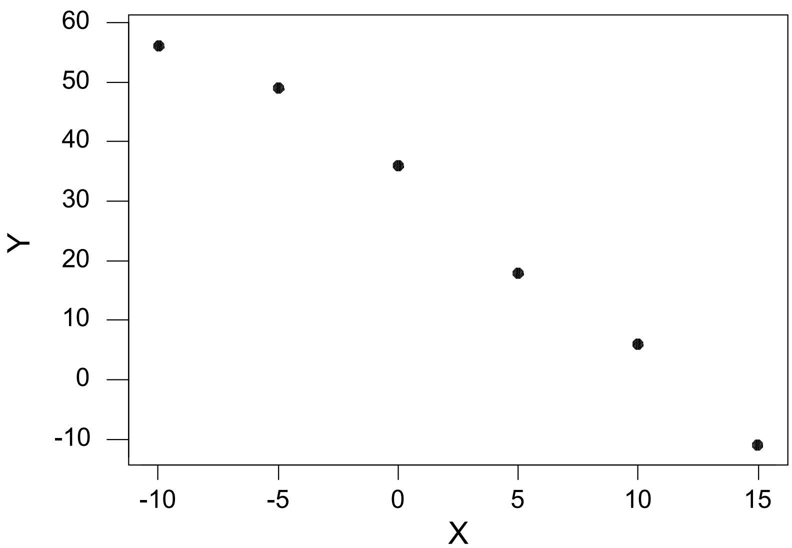
- RELATIONSHIP IS LINEAR (FOLLOWING A STRAIGHT LINE)
- RESIDUALS (RANDOM DEVIATIONS OF INDIVIDUAL *Y* VALUES FROM THIS STRAIGHT LINE) ARE NORMALLY DISTRIBUTED

SAMPLE HAS A FORM OF A TABLE, E.G.

X (°C)	-10	-5	0	5	10	15
Y ($)	56	49	36	18	6	-11

(*Y* IS DAILY NET PROFIT OF A TEENAGE ENTREPRENEUR SELLING HOT CHOCOLATE, *X* IS THE CORRESPONDING DAY'S AVERAGE TEMPERATURE).

THE DATA CAN BE DISPLAYED GRAPHICALLY IN SO CALLED SCATTER DIAGRAM:



TO ANSWER ALL POTENTIAL QUESTIONS, WE FIRST COMPUTE

Ex, Ey, Ex^2, Ey^2, Ex@y

THEN CONVERT THEM TO

$$\bar{x} = \frac{\Sigma x}{n} \qquad \bar{y} = \frac{\Sigma y}{n} \qquad SS_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$SS_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n} \qquad SP_{xy} = \Sigma x \cdot y - \frac{(\Sigma x)\cdot(\Sigma y)}{n}$$

HAVING COMPUTED THESE FIVE BASIC QUANTITIES, WE CAN NOW FIND THE REGRESSION COEFFICIENTS (INTERCEPT AND SLOPE) OF THE BEST (LEAST SQUARES, REGRESSION) STRAIGHT LINE:

$$b = \frac{SP_{xy}}{SS_x}$$ (SLOPE FIRST) $$a = \bar{y} - b \cdot \bar{x}$$

NOTE THAT THIS LINE MUST PASS THROUGH THE $(\bar{x}, \bar{y})$ POINT.

NEXT, WE NEED TO COMPUTE RESIDUAL STANDARD DEVIATION (TYPICAL MAGNITUDE OF THE RESIDUALS) BY:

$$s_r = \sqrt{\frac{SS_y - b \cdot SP_{xy}}{n-2}}$$

WHERE RESIDUALS ARE DEFINED AS

$$y_i - (a+b@x_i)$$

THE PREDICTION INTERVAL (ALSO CALLED CONFIDENCE INTERVAL FOR PREDICTION) FOR A NEW Y OBSERVATION, TAKEN AT $X = x_0$ (A SPECIFIC NUMBER) IS:

$$a+b\cdot x_0 \pm t_c \cdot s_r \cdot \sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{SS_x}}$$

WHERE t_c IS THE CORRESPONDING CRITICAL VALUE OF THE STUDENT DISTRIBUTION, USING $n - 2$ DEGREES OF FREEDOM.

SAMPLE CORRELATION COEFFICIENT IS COMPUTED BY

$$r \equiv \frac{SP_{xy}}{\sqrt{SS_x \cdot SS_y}}$$

(ALWAYS BETWEEN -1 AND 1).

COEFFICIENT OF DETERMINATION: r^2
(PERCENTAGE REDUCTION OF THE ORIGINAL *y* VARIANCE).

TWO TESTS

H_0: $\$ = 0$ .... WHERE $\$$ IS THE SLOPE OF THE POPULATION STRAIGHT LINE

TEST STATISTIC: $\dfrac{b}{s_r} \cdot \sqrt{SS_x}$ (USE t_{n-2})

H_0: $D = 0$ .... WHERE D IS THE POPULATION CORRELATION COEFFICIENT

TEST STATISTIC: $\dfrac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$ (USE t_{n-2})

ONE CAN SHOW THAT THE TWO TEST STATISTICS (THUS THE TWO TESTS) ARE IDENTICAL.